**M8 Portfolio Project Option 1**

Krista O’Neill

Colorado State University Global

MIS543 M8 Portfolio Project Option 1

Dr. Shawn Harrs

March 13, 2022

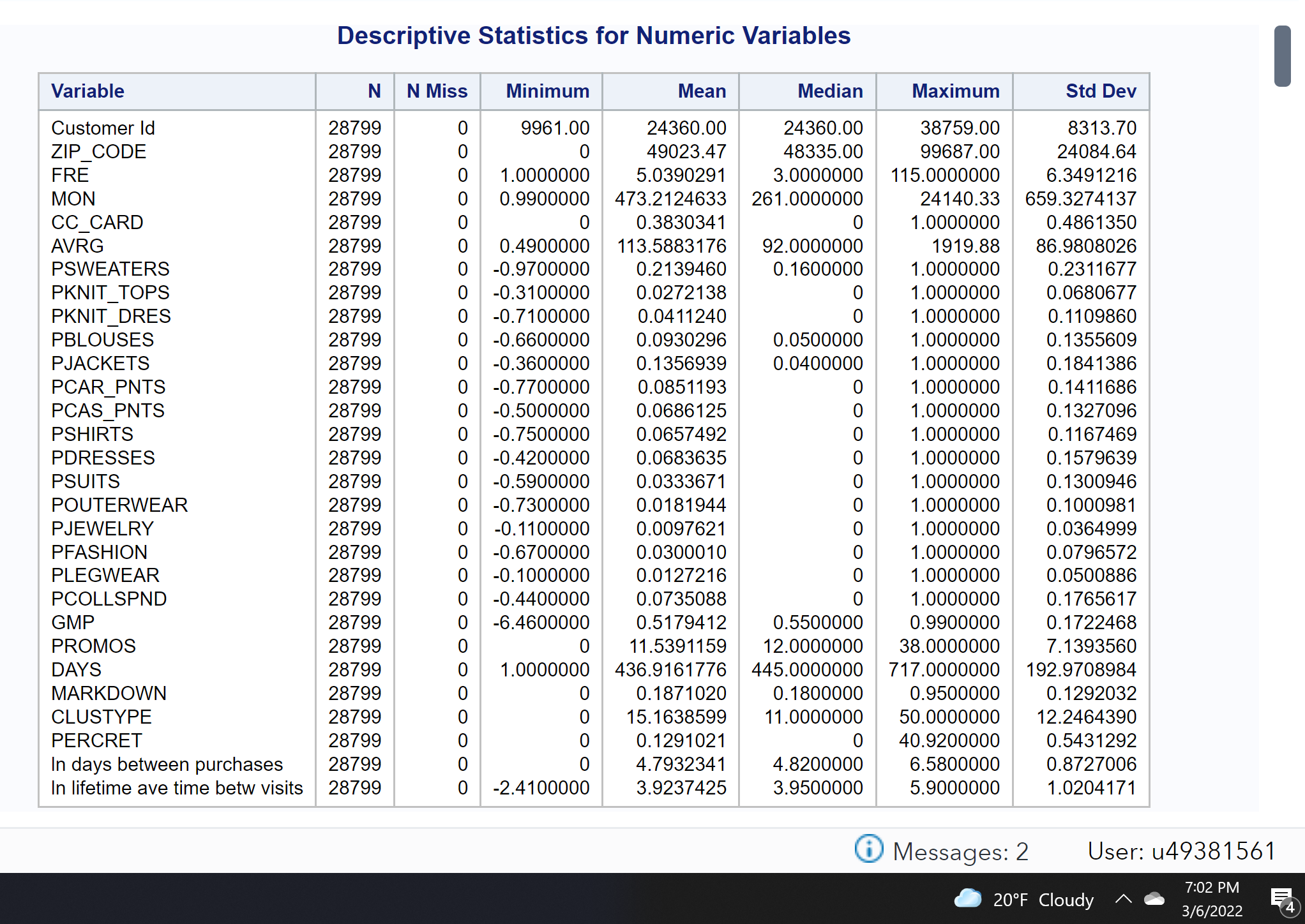
**Introduction**

A retail clothing store is seeking to increase their sales while marketing more effectively. Their primary strategic goals are to identify which customers respond to marketing promotions. Secondarily, they want to be able to predict future growth. By a careful analysis of the dataset, an analyst should be able to create, test/model, and answer targeted business questions efficiently. These business questions based on the organization’s strategy will form the basis of a complete portfolio project.

**Data Overview**

**Figure 1**

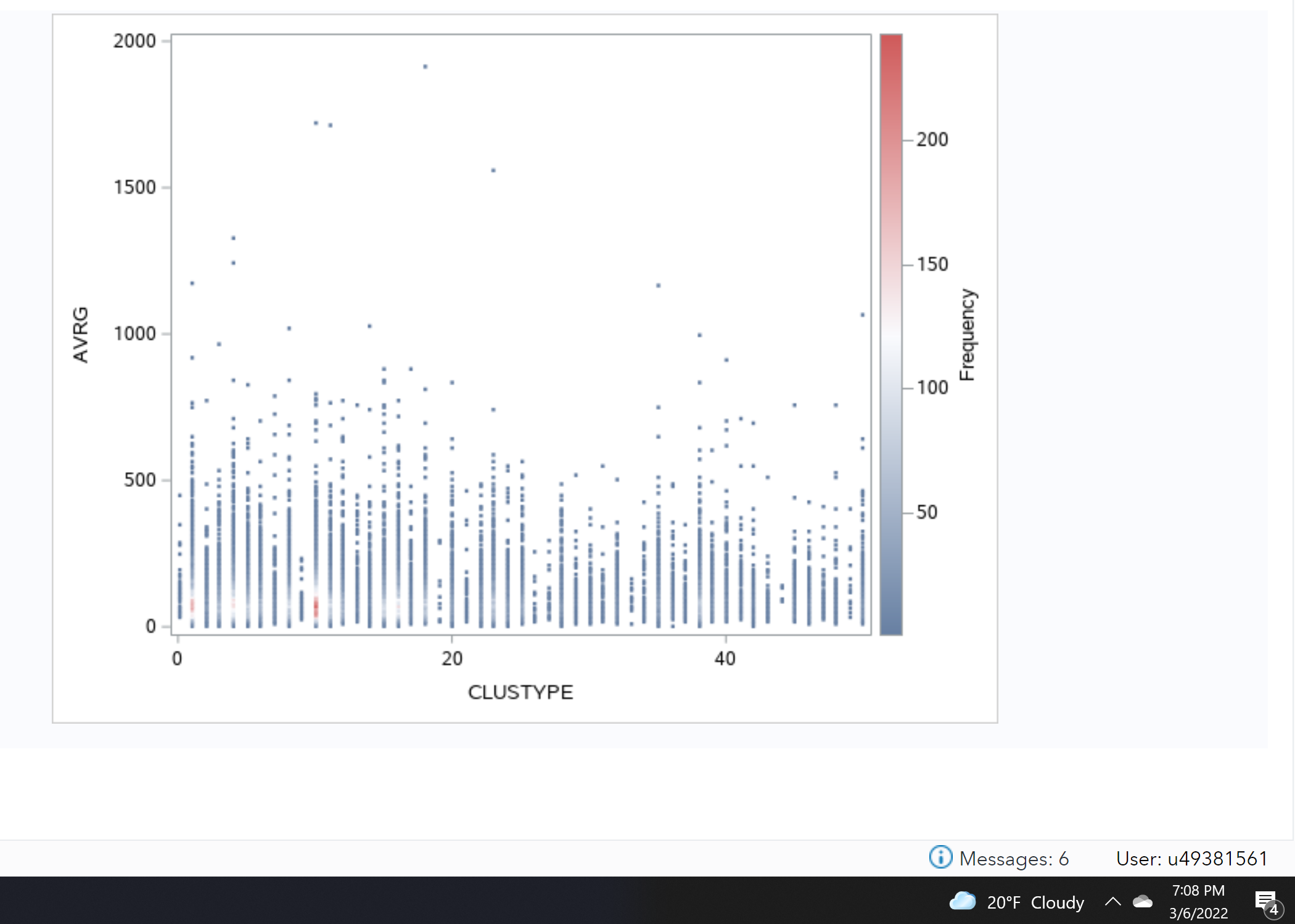
*Overview of descriptive statistics for entire dataset*



There are no missing values in the dataset, which will eliminate the difficulties this can cause. It must be noticed that this is a numerical statistical summary, meaning that variables with meaningless numerical values (like customerID or ZIP\_CODE), do not give statistical information. A quick look, though, shows that variables like AVRG (average) and DAYS show high standard deviations, meaning there could be further investigation into why there is a larger variance in the average than the other variables.

**Figure 2**

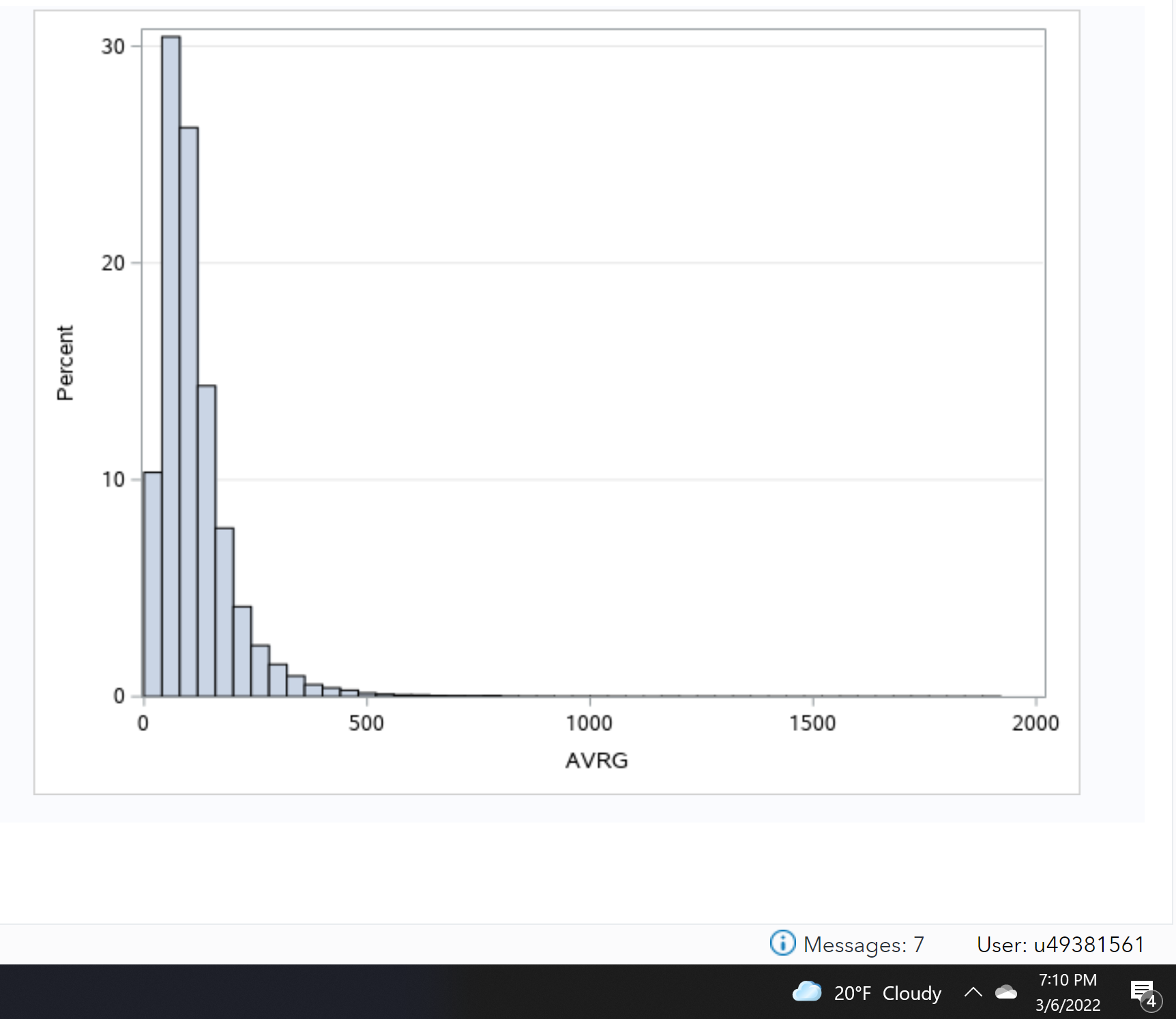
*Showing a heatmap of the average spent based on cluster type.*



A quick analysis shows that Cluster 112 has a higher frequency of spending, even though that frequency is at a lower average than others. Cluster 20 has an outlier with the highest spending. This simple heatmap can give the company a very quick overview of the spending habits of each cluster on average; an analyst can look at this graph and see that Cluster 26 and 27 spend very little in comparison to Cluster 5 and Cluster 2.

**Figure 3**

*Showing a histogram of the average spending on whole*



On average, spending is skewed right, with far more spending between $0-500 than any higher than that.

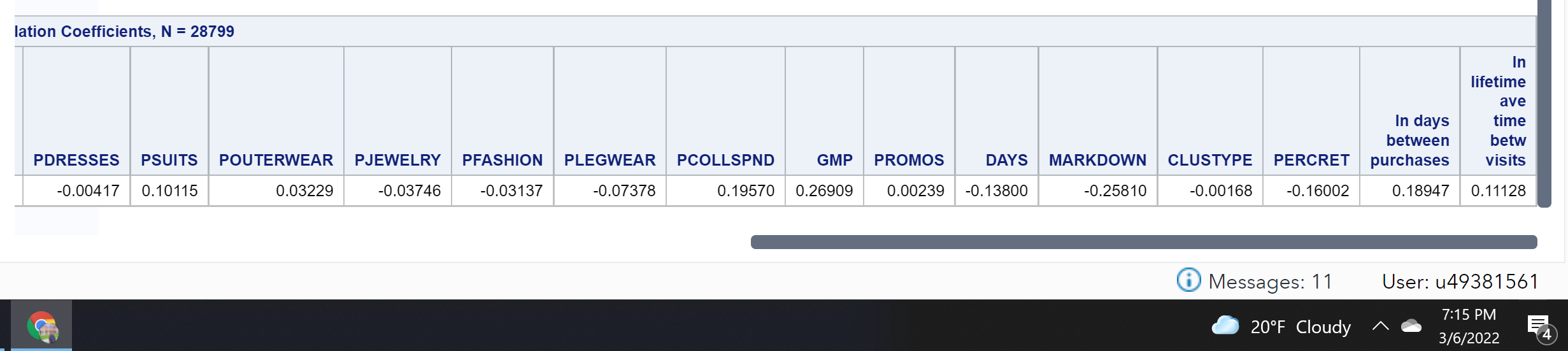
**Figure 4**

*Showing correlations of every other variable, with the variable “AVERAGE”*



**Figure 5**

*Showing the correlation, continued*



**Business Questions, Hypothesis, & Test 1**

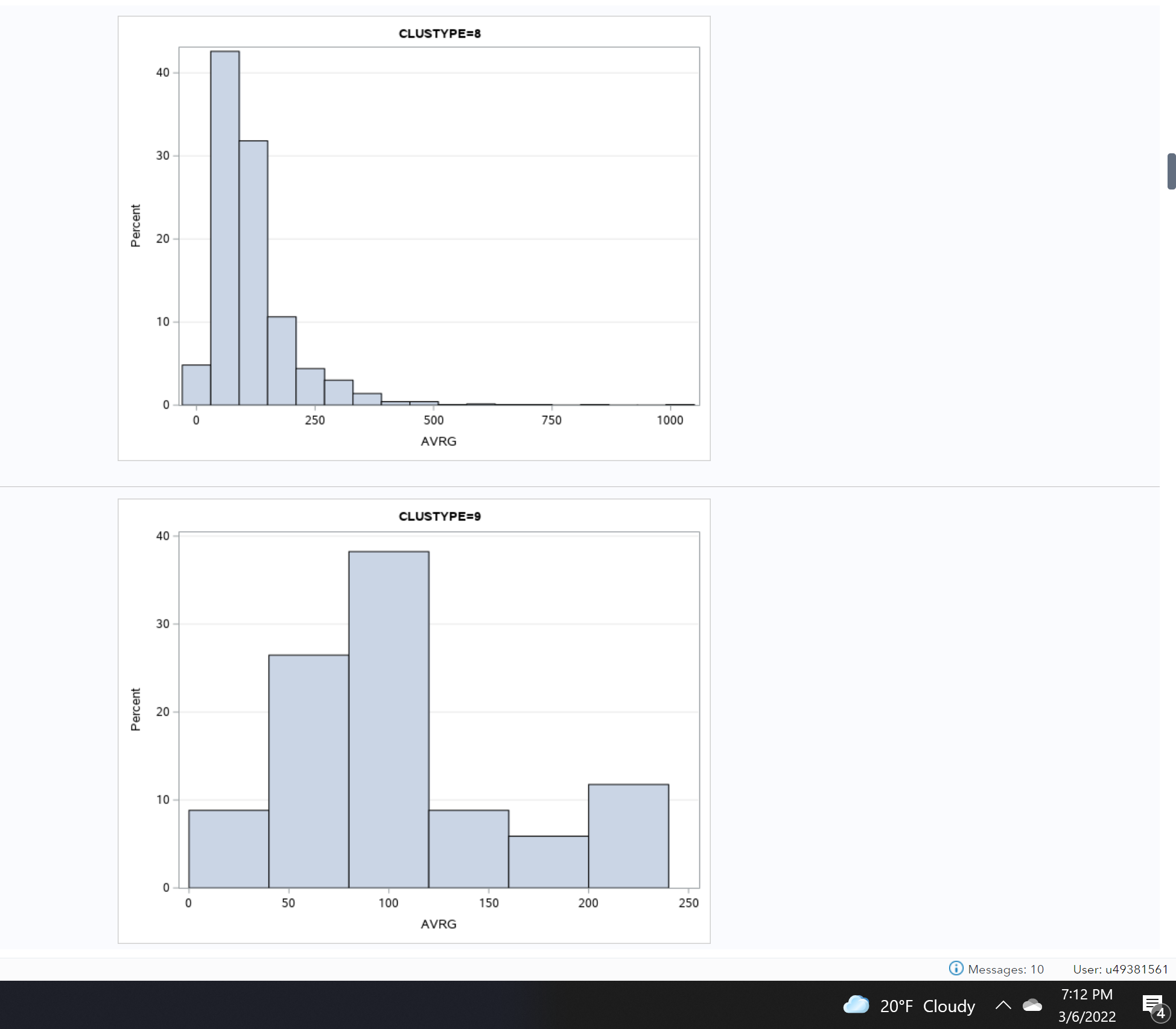
The organization is seeking to identify customers who respond to marketing. The first business question to answer is then, *Do some clusters of customers respond to marketing more than others?* By identifying the cluster of customers most responsive to marketing efforts, the organization then will know which clusters it should target when expanding its marketing campaigns, improving overall advertising engagement.

H0: *The cluster of customer does not impact engagement with marketing.*

H1: *The cluster of customer does impact engagement with marketing*.

**Figure 6**

*Showing a grouped histogram of average spending*



When grouping histograms of average spending by cluster type, it becomes evident that some clusters spend, on average, a higher amount. As pictured in Figure \_\_\_, Cluster 9 shows a far less skewed distribution of average spending.

Furthermore, a one-way ANOVA test with AVRG as the dependent variable and CLUSTYPE as the categorical variable showed the p-value is 0.0028, or significantly less than 0.05. This indicates that the cluster type is statistically significant, and therefore the null hypothesis can be rejected. The cluster type does have an impact on overall average spending.

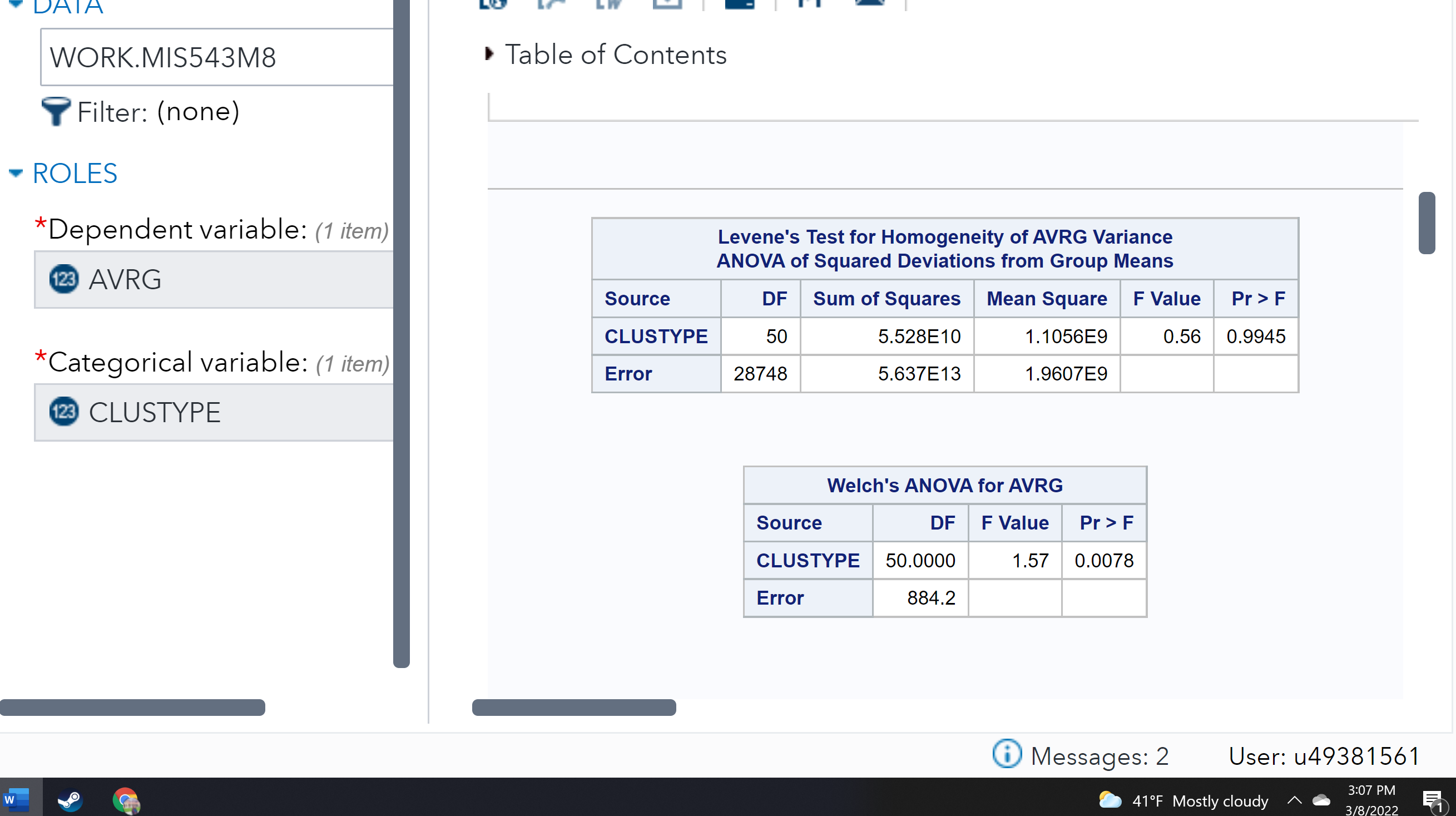
**Figure 7**

*Showing the one-way ANOVA test*

The Levee’s test for Homogeneity has a p-value of much higher than 0.05, at 0.9945, suggesting that due to the low p-value for Welch’s test, homogeneity can be rejected; the clusters are not the same.

**Figure 8**

*Showing the Levene’s vs. Welch’s test*



Finally, in order to correlate the marketing promotions with average spending and cluster, a correlation analysis using “PROMOS” correlated with “AVRG” and “CLUSTYPE” shows very low correlation coefficients.

**Figure 9**

*Showing the correlation analysis*

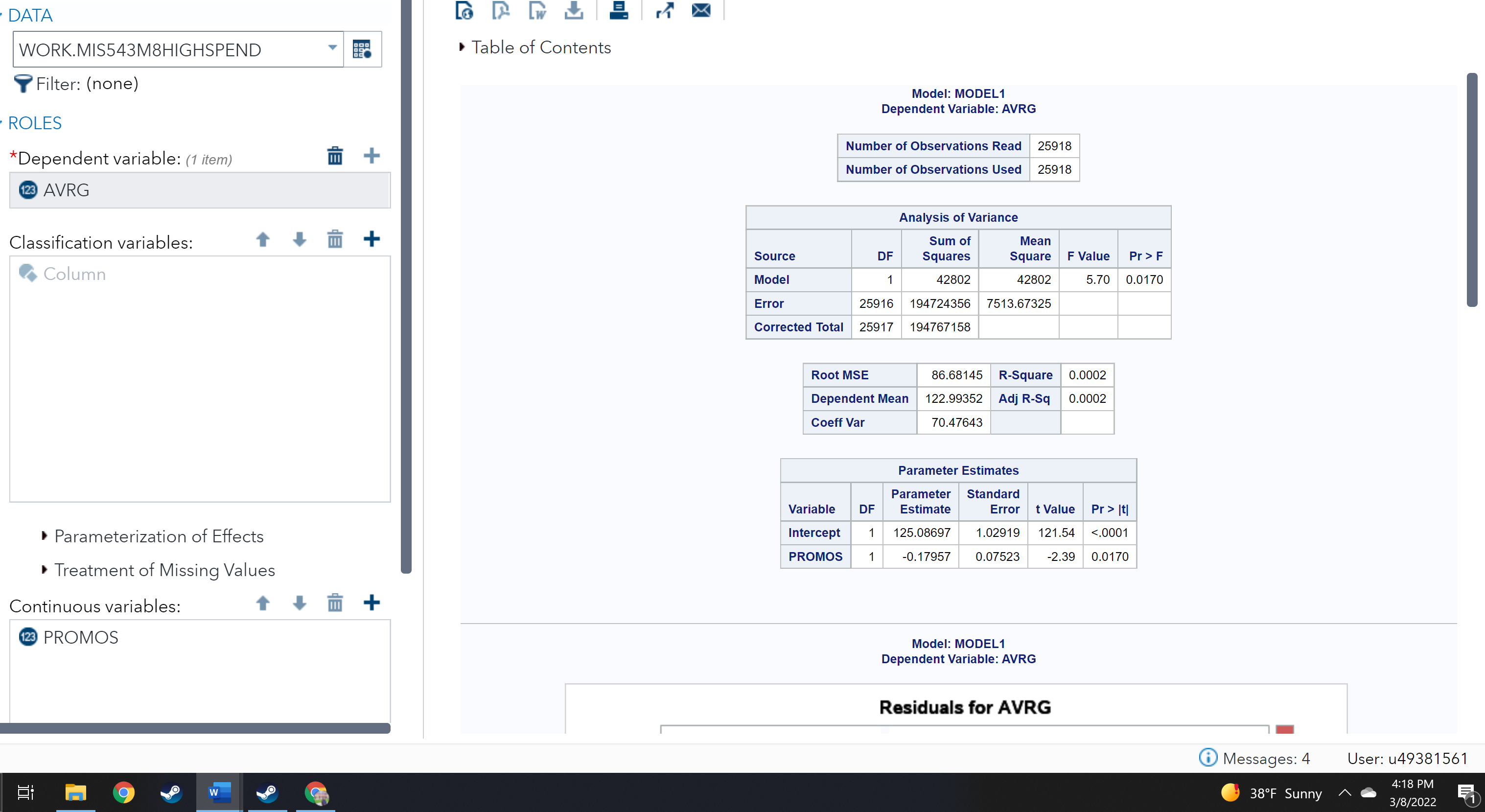


This suggests that in while the cluster is related to average spending, the amount of promotions does not correlate with either one.

Finally, the data was filtered in order to show only observations where AVRG spending was higher than the 10th percentile. Then a linear correlation was done on this filtered dataset using PROMOS as a continuous variable, AVRG as the dependent variable.

**Figure 10**

*The results of the linear correlation with PROMOS as a continuous variable, AVRG dependent*



While the p-value is less than 0.05, indicating statistical significance, the correlation coefficient for PROMOS indicates a negative value of -0.17957. This means that the more promotions the high-spending customer received, the lower their average spending. This means that the null hypothesis must be rejected and the alternative accepted. More promotions on file for high-spending customers decreased the average amount of spending.

**Business Question, Hypothesis, & Test 2**

The organization also would like to increase its sales. It might be able to do so by addressing individual factors that might impact customer experience. As technology moves forward, businesses should consider alternate methods of payment, such as cash, Google Pay, or other wireless/cardless options. A business question they might answer is, *Does the use of a credit card affect sales?* This question can be answered with a two-sample t-test. Should the alternate hypothesis prove true, the store should consider ensuring efficient and convenient methods of alternate payment.

H0: *Sale amount is not correlated to the usage of a credit card*.

H1:*Sale amount is correlated to the usage of a credit card.*

**Figure 11**

*A two-sample t-test showing using CC\_Card as the groups variable, AVRG as the analysis variable*.



Because the equality of variances value is less than 0.05, it can be assumed that the two sample variances are not equal. Therefore, the Satterthwaite “unequal” p-value should be used. Because that p-value is <0.001, it can be concluded that the usage of a credit card does in fact impact average sales. The null hypothesis can be rejected, and the alternative accepted. Looking at the means of the values provided above, it can be seen that customers who used a credit card spent a statistically significantly larger amount than customers who did not.

**Business Question, Hypothesis, & Test 3**

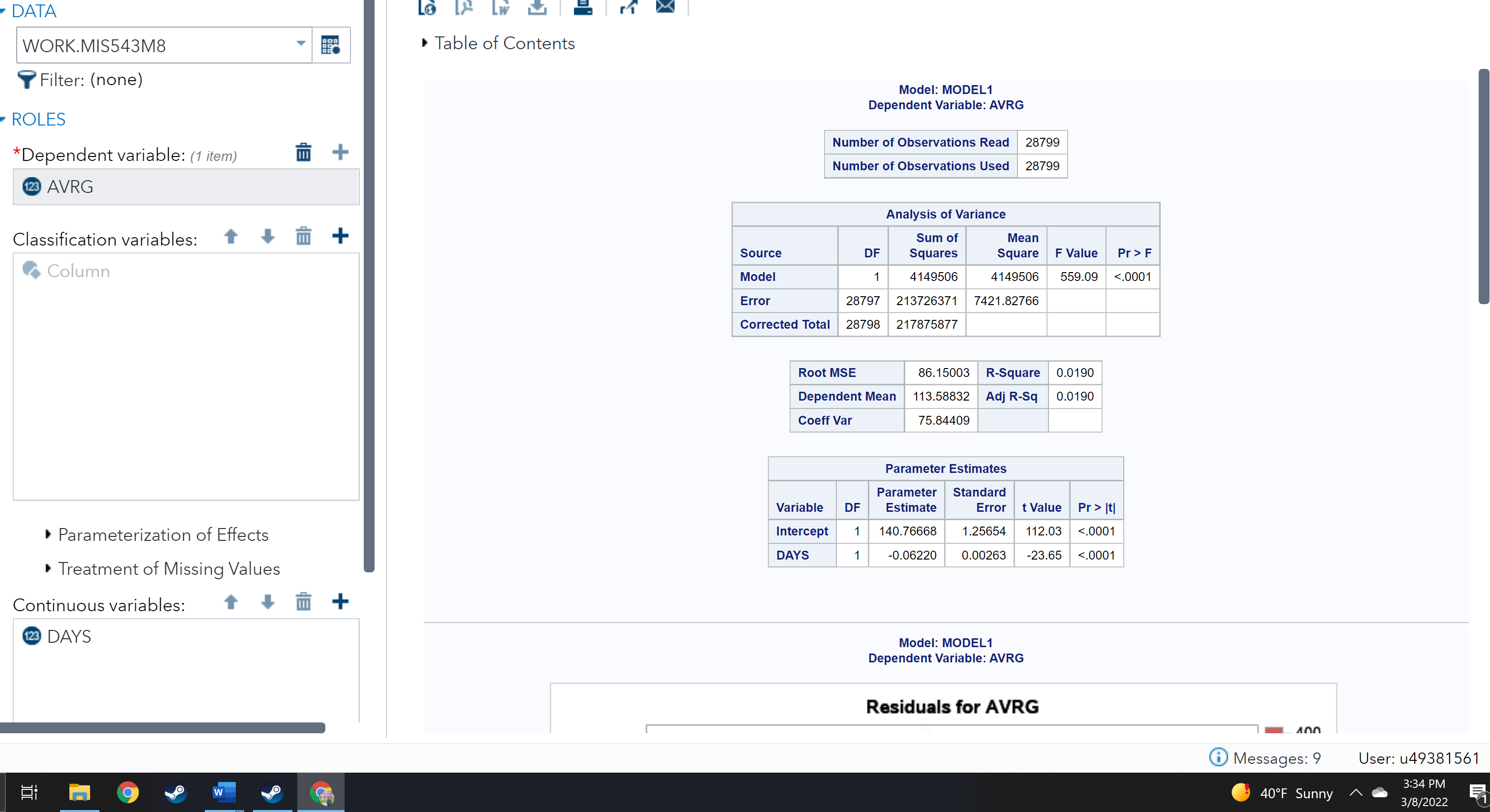
The business also wants to be able to predict future business growth. One way of measuring that might be to answer the business question, *Do our customers spend consistently more, the longer they are members of our organization?* If not, then the business might want to dig deeper into customer data and evaluate why customers are not spending more. The same question could be asked again, but in regards to an increase in visits over membership time. If increased membership time increases sales, the business could use a regression to estimate monthly net sales in the future.

H0: *Increased membership time does not correlate with net sales.*

H1: *Increased membership time does correlate with net sales*.

**Figure 12**

*Showing a linear regression using DAYS as a continuous variable, AVRG as the dependent.*



The linear regression shows a p-value of far less than 0.05, showing that it is statistically significant. However, the coefficient of the variable DAYS is a negative value of -0.06220, indicating that in fact, the average amount spent by a customer decreases the longer that they are on file. The null hypothesis must be rejected, and the alternative hypothesis accepted. The business loses money the longer a customer is on file. A similar test using GMP (gross margin percentage) showed the same results, a longer membership correlated with lower earnings for the company.

**Figure 13**

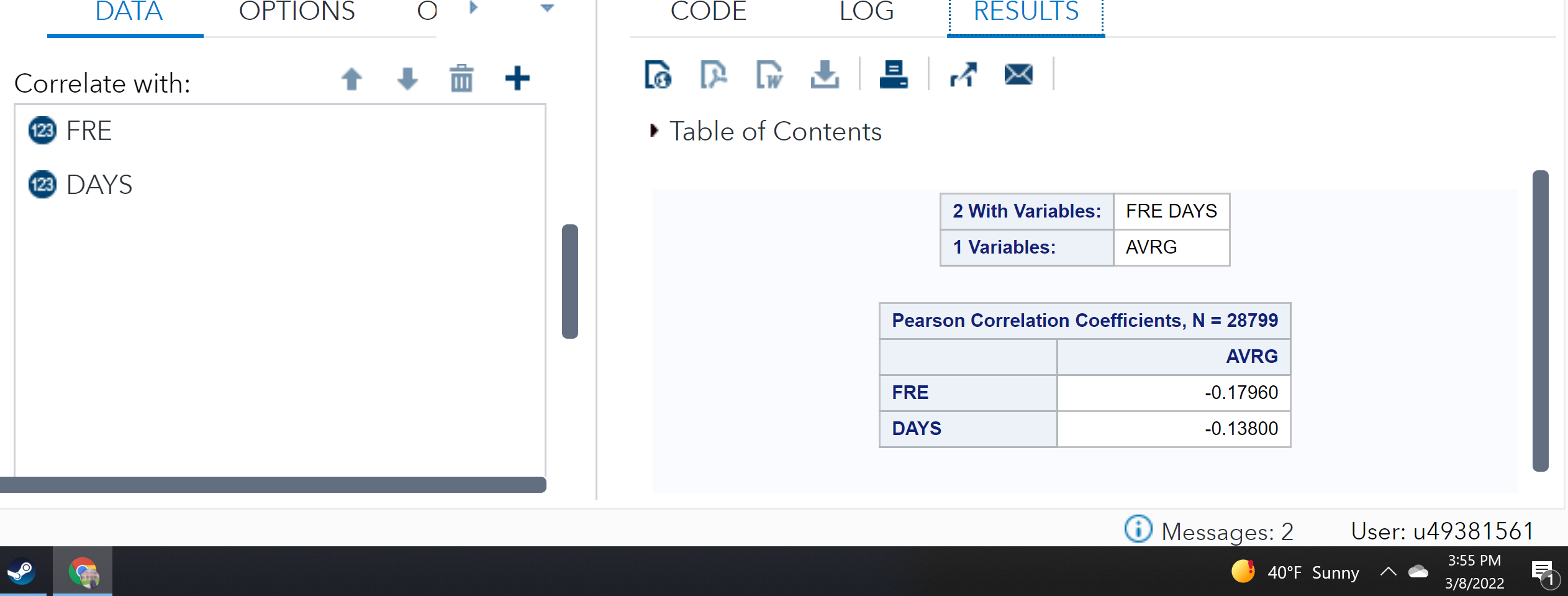
*A similar linear regression using GMP as the continuous variable instead*.



An additional correlation analysis between AVRG using the variables DAYS and FREQ also reveals negative correlations, indicating that the longer a customer is on file and the more often they visit, the less they spend.

**Figure 14**

*Showing the correlation between AVRG, and Days and FREQ*



**Business Question, Hypothesis, & Test 4**

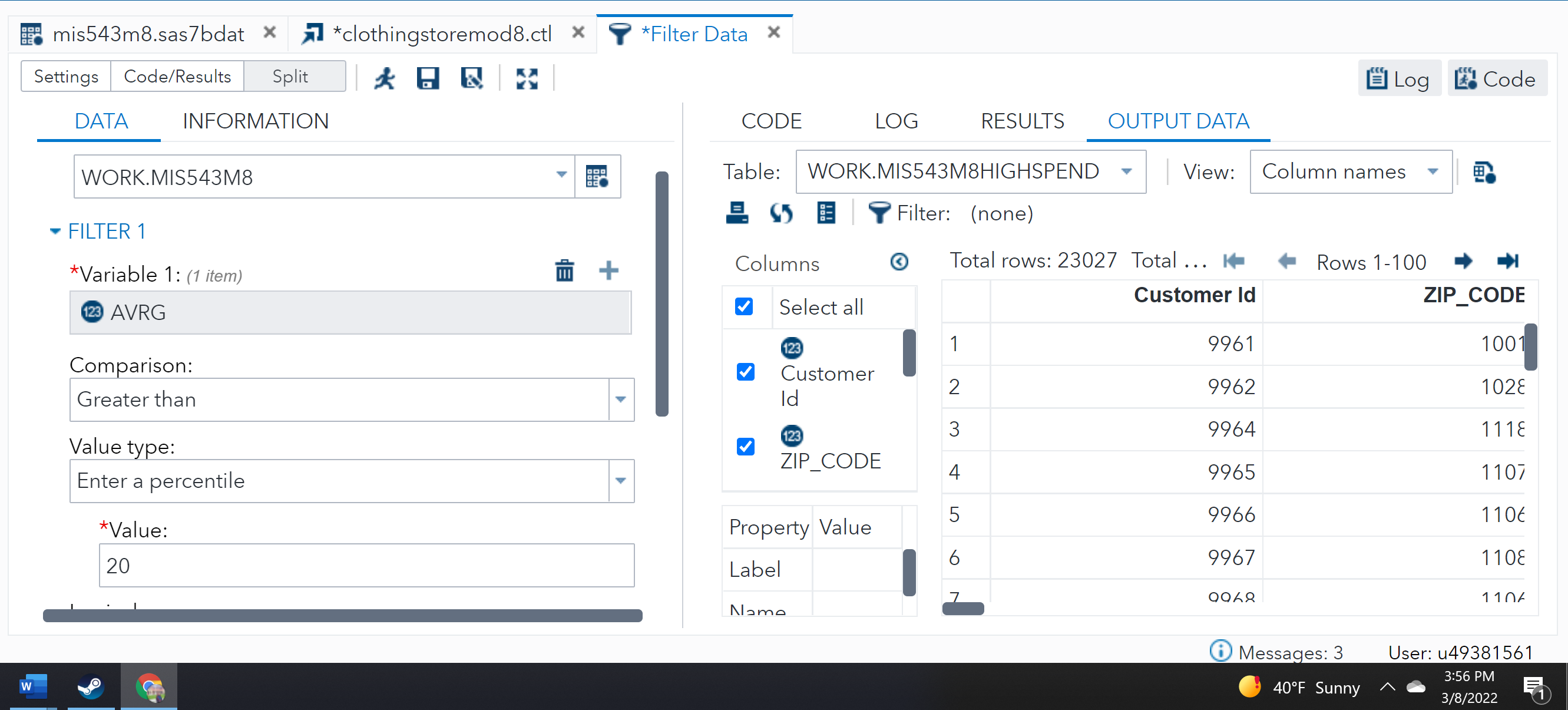
Finally, the business wishes to maximize sales. This could be done by finding out the cluster of customers who spend the most on average, and then seeing which category of items they spend on consistently more. The business question would be, *Which items are in high demand from high-spending customers?* Should they discover that specific categories are in demand, then they could consider prioritizing a higher variety of said items; conversely, they could estimate how much they might be losing in profits by retailing the low-selling items, or how to market them and sell them more effectively.

H0: *Product category is not correlated to average amount spent per visit.*

H1: *Product category is correlated to average amount spent per visit*.

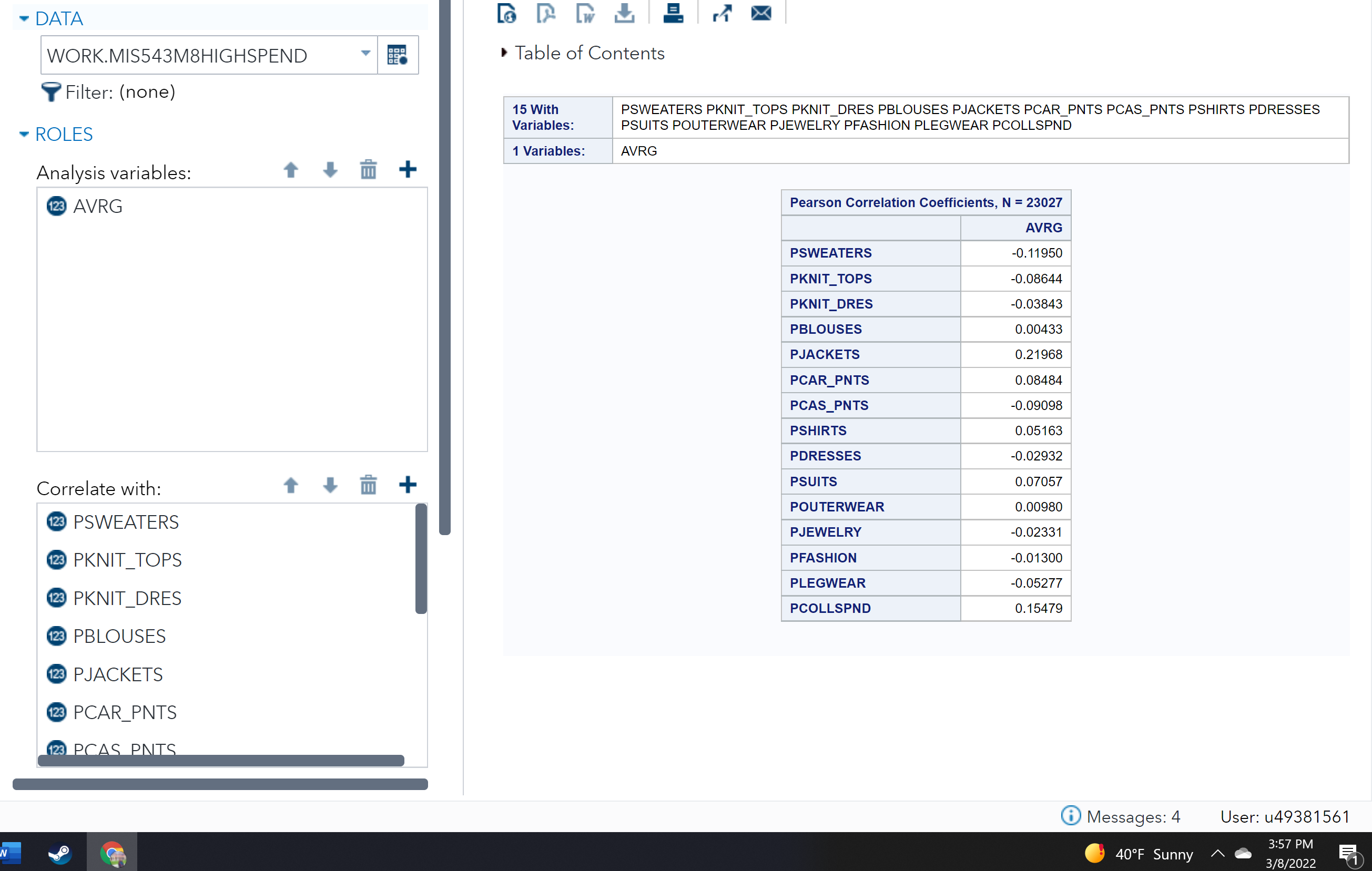
**Figure 15**

*Filtering data by those whose AVERG spending is higher than the 20th percentile*



**Figure 16**

*Showing a correlation analysis of each product category in relation to AVRG*

******

The correlation analysis showed negative values for several of the categories, such as sweaters, knit tops, and casual pants. The negative correlations for these variables are higher than those of jewelry or legwear, indicating that they sell less well. Jackets have a relatively high positive correlation factor of 0.21968, as do collectibles at 0.15479. However, none of the correlation values are close to -1 or 1, indicating weak correlations. Because none of these values are higher than even 0.5, it can be concluded that the alternative hypothesis must be rejected and the null hypothesis accepted. The category of item sold does not truly affect average spent per customer, even among the high-spending customers.

**Recommendations**

With respect to business question 1, regarding the response of customers to marketing, it appears that the company’s marketing promotions are not having the desired effect on any customers, especially the ones that are spending a higher amount than others. This may be a result of customers being bombarded with promotions they do not know how to take advantage of (Hu et al., 2018). The company should investigate reasons the promotions are not effective. A strategy they could consider is attempting to tailor marketing more specifically to clusters, so that the promotions more specifically target their audience. By reducing quantity and adding quality, the company could make the promotions more helpful and less overwhelming (ibid). They should also consider collecting subjective customer feedback on the promotions being offered, in order to gain valuable further perspective on the reasons increasing promotions decreases average sales.

Business question 2 focuses on the usage of credit cards. The statistical results showed that customers who paid with a credit card were more likely to spend more money. As technology moves forward, the company should consider preemptively investing in providing alternate methods of payment, such as cash, Google Pay, or other wireless/cardless options. An investigation of the demographics of customers who are paying without credit cards would also benefit the company. If this is an issue of not having their preferred method of payment available, or inconvenience, the company could seek to address this. If it is a matter of demographically being less likely to have the finances to buy things at the store, the company could consider providing promotions that would increase the GMP of those customers. Additionally, by examining the addresses and physical location of customers more likely to have a credit card, the store can explore options about where would be geographically best to open new locations (Pahuja et al., 2021).

Business question 3 revealed that the longer a customer is on file, the lower their average spending per visit. While this could be a simple matter of the customer not needing to purchase anything, it also could signify a lack of customer satisfaction with the store experience. There are many factors that go into a customer’s likelihood to churn, especially when a retail experience involves many factors like ease of transportation to the location, physical environment, and space, as well as price and demand (ibid). In order to explore this data further, the company could attempt to identify if there is any particular store or location that is more likely to suffer sales degradation over time. By performing a comparison between locations, they might be able to pinpoint reasons why customers spend less the longer they are a member. The same could be said for analyzing the demographics and lifestyles of customers who have a sharp falloff rate.

Business question 4 showed that there was no particular draw to any of their categories of products, based on those customers who spent the most on average. Depending on the type of store they are, the company could consider specializing, or dividing their company into multiple brands with different focuses of sale. This goes back to the company’s mission or goal. If they are a department store, then their goal is to provide a wide variety of items. However, providing a specialized experience might draw customer attention, as well as improve the customer experience due to a more knowledgeable staff (Hartman, 2019). Customers are often more willing to pay for an experience that is branded as more luxury, tailored, and specific to individual needs (Rombach et al., 2018). This is a potential avenue the company could explore.

**Conclusion**

This retail store has collected a significant amount of data that will help them explore their business options. While only four business questions were explored, there is much more that could be asked and answered from this dataset. With just a brief analysis, the company has been able to see that they have room for improvement. By collecting further data, especially regarding customer cluster demographics, and exploring further models, visualizations, and algorithms, they will be able to widen their perspective and understand how better to attract and keep their customers.

**References**

Aversa, J., Hernandez, T., & Doherty, S. (2021). Incorporating big data within retail organizations: A case study approach. *Journal of Retailing and Consumer Services* 60. <https://www-sciencedirect-com.csuglobal.idm.oclc.org/science/article/pii/S0969698921000138>

Grigsby, M. (2016). Advanced customer analytics: Targeting, valuing, segmenting and loyalty techniques. *Business & Economics*. <https://csuglobal.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=e020mna&AN=1365518&site=eds-live&ebv=EB&ppid=pp_i>

Hartman, D. (2019). Advantages and disadvantages of a specialty store. *Chron*. <https://smallbusiness.chron.com/advantages-disadvantages-specialty-store-22760.html>

Hu, W., Tang, S., Chen, Y., Yu, C., & Hsu, W. (2018). Promotion recommendation method and system based on random forest. *MISNC '18: Proceedings of the 5th Multidisciplinary International Social Networks Conference, 1-5.* <https://dl-acm-org.csuglobal.idm.oclc.org/doi/10.1145/3227696.3227708>

Pahuja, H., Delir, P., Li, Y., & Jayaraman, P. (2021). *The 23rd International Conference on Information Integration and Web Intelligence*, 49-56. <https://doi-org.csuglobal.idm.oclc.org/10.1145/3487664.3487672>

Purdue (n.d.). *One way ANOVA (or Welch’s test)*. <https://www.stat.purdue.edu/~tqin/system101/method/method_one_way_ANOVA_sas.htm>

Rombach, Mm., Widmar, N., Byrd, E., & Bitsch, V. (2018). Do all roses smell equally sweet? Willingness to pay for flower attributes in specialized retail settings by German consumers. *Journal of Retailing and Consumer Services 40, 91-99.* <https://www-sciencedirect-com.csuglobal.idm.oclc.org/science/article/pii/S0969698917300103>